Evaluating Cost-Accuracy Trade-offs in Multimodal Search Relevance Judgements*

Silvia Terragni†, Hoang Cuong, Joachim Daiber, Pallavi Gudipati and Pablo N. Mendes*

Objective, Inc. San Francisco, CA, USA.

Abstract

Large Language Models (LLMs) have demonstrated potential as effective search relevance evaluators. However, there is a lack of comprehensive guidance on which models consistently perform optimally across various contexts or within specific use cases. We argue that the evaluation of model performance is inherently complex due to the numerous variables involved in the process. We explore the trade-off between cost and accuracy by guiding smaller LLMs to attain a level of quality comparable to that of larger commercial LLMs, while significantly reducing associated costs.

Keywords

Multimodal Search, Relevance Judgments, Large Language Models, Multimodal Large Language Models

1. Introduction

Search relevance evaluation is the process of assessing how effectively an information retrieval system returns results that are relevant to a user's search query. The process typically involves multiple human judges, tasked with stating whether each search result is relevant to a search query. The resulting relevance judgements are then aggregated through evaluation metrics to quantify relevance. Those in turn enable researchers and practitioners to compare different retrieval systems in order to select the best option for a given application.

Multimodal Search presents additional challenges in search relevance evaluations due to the complexity of interpreting and integrating information from various attributes across different modalities. For instance, in e-commerce search, assessing relevance requires understanding the intent behind the search query and comparing it with a judge's interpretation of product relevance based on multiple features including the title, description, and images, as well as other attributes such as category, color, and price. The task is further complicated by different characteristics across use cases. For instance, when searching for very visual aspects (e.g. design assets) the images play a much more central role, as compared to other use cases where product category and other attributes are more important (e.g. searching for hotel supplies). Data quality also varies significantly by use case. In applications with user-generated content, data may be missing or low quality – e.g. product descriptions often conflict with the information that can be gleaned from images.

While human annotators remain the most reliable source for obtaining relevance judgments, the process is costly, and time-consuming. Recent work [1][2] has shown that Large Language Models (LLMs) and Multimodal Language Models (MMLMs) are a viable alternative to produce relevance judgements. LLMs-as-judges are enticing since they can unlock higher relevance judgement throughput at a fraction of the cost. As a result, they offer the potential of widespread relevance improvement in search systems due to more accessible and extensive evaluations, as well as training data generation. However, progress is hampered by a number of under-explored questions about how to best employ LLMs-as-judges.

MMSR'24

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

In this paper we evaluate a number of LLMs and MMLMs in terms of their alignment with human judgements and ask the following research questions:

1. Is LLM performance use-case dependent? In other words, would the same LLM perform well in one use case but not in another?
2. Is there a clear winner? In other words, is there a model that consistently outperforms all the others across all use cases?
3. Is multimodal support necessary for search relevance judgement in multimodal search?
4. What models offer the optimal cost-accuracy trade-offs?

In the next section we summarize related work. We then present our experimental setting, and discuss results. Finally, we present concluding remarks and future work.

2. Related Work

Large Language Models (LLMs) have shown exceptional abilities in a wide variety of tasks, and using them for evaluating Information Retrieval systems is receiving considerable attention [1]. Recent studies have explored different methods for generating relevance judgements. For example, Prometheus [3] is a 13-*billion* parameter LLM designed to evaluate long texts using customized scoring rubrics provided by users. JudgeLM [4] uses fine-tuned LLMs as scalable judges to evaluate other LLMs effectively in open-ended tasks. They find that JudgeLM has high agreement with expert judges, over 90%, and works well in evaluating single answers, multimodal models, multiple answers, and multi-turn dialogues. [5] develops an LLM prompt based on feedback from search engine users. They show accuracy similar to human judges and can identify difficult queries, best results, and effective groupings. They also find that both changes to prompts and simple paraphrases can improve accuracy.

In the context of Multimodal LLMs (MMLMs), Chen et al. [2] assess these models as judges through a new benchmark. They examine their performance in tasks such as Scoring Evaluation, Pair Comparison, and Batch Ranking. The study points out that MLLMs need more improvements and research before they can be fully trusted, as they can have biases like ego-centric bias, position bias, length bias, and hallucinations. Additionally, [6] investigates the relevance estimation of Vision-Language Models (VLMs), including CLIP, LLaVA, and GPT-4V, within a large-scale ad hoc zero-shot retrieval task aimed at multimedia content creation.

To the extent of our knowledge we are the first to compare the cost-accuracy trade-offs of several generally available LLMs of different sizes.

3. Methodology

This study employs a relevance evaluation process to assess the performance of LLMs and MMLMs (collectively referred to as “models”) for search relevance judgements. We assess these models based on two critical dimensions: accuracy and costs. Our evaluation pipeline consists of three stages:

- *Data Collection*: We obtained search results from three datasets across different domains using a list of predefined queries.
- *Human Annotation*: Two trained human annotators assigned relevance grades to each (query, result) pair following some established relevance criteria.
- *Model Evaluation*: We applied a range of LLMs and MLLMs to generate relevance judgments for the same sets of search results, comparing their performance against human annotations.

Each stage is discussed in detail in the following subsections, where we describe the datasets, retrieval system, grading strategy, and the models used.

3.1. Datasets

We conducted our experiments on three datasets: *Fashion*, *Hotel Supplies*, and *Design*. The Fashion dataset is a subset of the publicly available dataset *H&M Personalized Fashion Recommendations*.[1] The Hotel Supplies and Design datasets are proprietary and represent domains in e-commerce search for hotel supply products, and social media search for design assets, respectively. Each dataset includes multiple textual fields per product, along with one or more associated images. Table 3.1 summarizes the characteristics of each dataset, detailing the average number of fields per search result, the average number of empty fields, and the average word count per result. These factors can impact the difficulty of generating relevance judgments.

Dataset	Total Number of Search Results	Avg Number of Textual Fields	Avg Number of Empty Textual Fields	Avg Number of Words per Result
Fashion	1120	33	1	49
Hotel Supplies	2210	17	8	96
Design	1713	32	3	69

Table 1
Summary statistics of the used datasets.

3.2. Retrieval System

To obtain relevant search results, we utilized a baseline retrieval system that combines BM25 with one of the top-ranked text embedding models models [7] in the MTEB Leaderboard [8] as of June 2024. We created indexes for each dataset to enable efficient retrieval of results based on predefined list of queries. These queries were either derived from real traffic data or carefully crafted by human experts to ensure they represented a wide range of search scenarios. Our aim was to include queries and results that included hits and misses generated by both lexical and semantic retrievers.

3.3. Relevance Judgement Strategy

Each dataset was structured as a collection of query-result pairs. Two expert human annotators assessed the relevance of each pair on a 0-2 rating scale:

- 2: Highly relevant, a perfect match for the query;
- 1: Somewhat relevant, a result that partially matches the query's intent.
- 0: Not relevant, a poor result that should not be shown.

The human annotators were provided with detailed guidelines to ensure consistency in their relevance judgments. Table 3.3 provides examples of different relevance judgment categories for the query "*v-neck white tee*". In the first row, the result is highly relevant, as both the image and the text describe a white v-neck t-shirt. Therefore the human relevance judgment for this pair is a *2*. The second row shows a partial match: while the text mentions a white t-shirt, the image depicts a white v-neck t-shirt with black stripes, resulting in a relevance judgment of 1. The third row illustrates an irrelevant result (0), where the product shown is a strap top, unrelated to the query.

3.4. Inter-Annotator Agreement

To assess the reliability of the relevance judgments, either human or LLM-generated, we followed common practice and calculated *Cohen's Kappa* coefficient. Cohen's Kappa is a robust statistical measure

[1] https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations. The annotations for this dataset will be published upon acceptance.

Image	Search Result	Relevance Judgment
	prod_name: Premium ELKE vneck tee, index_name: Ladieswear, detail_desc: V-neck T-shirt in airy slub lin[...], department_name: Jersey/Knitwear Premium, index_group_name: Ladieswear, colour_group_name: White, product_type_name: T-shirt, graphical_appearance_name: Solid, perceived_colour_value_name: Light, perceived_colour_master_name: White	2
	prod_name: ED Lizzie tee, index_name: Ladieswear, detail_desc: Short-sleeved top in lightwei[...], department_name: Jersey, index_group_name: Ladieswear, colour_group_name: White, product_type_name: T-shirt, graphical_appearance_name: Stripe, perceived_colour_value_name: Light, perceived_colour_master_name: White	1
	prod_name: V-neck Strap Top., index_name: Ladieswear, detail_desc: V-neck top in soft organic [...], department_name: Jersey Basic, index_group_name: Ladieswear, colour_group_name: White, product_type_name: Vest top, graphical_appearance_name: Solid, perceived_colour_value_name: Light, perceived_colour_master_name: White	0

Table 2
Examples of three relevance judgment categories for the query "*v-neck white tee*", accompanied by corresponding search results. The descriptions of the search results have been shortened for brevity.

commonly employed to quantify inter-annotator agreement for categorical data. Cohen's Kappa values range from -1 to 1, where 1 indicates strong agreement, while values closer to 0 suggest agreement no better than chance. To interpret the Kappa values, we use the guidelines reported in Table 3.

In our evaluation, we compute Cohen's Kappa to measure the agreement between human annotators and LLM-generated annotations, as well as between pairs of human annotators. The degree of agreement between human annotators also provided insights into the difficulty of evaluating certain datasets.

3.5. Models

Our evaluation included a range of LLMs and MLLMs to reflect varying levels of performance and cost. We considered both large-scale proprietary models and more cost-efficient alternatives:

- OpenAI Models[2]: GPT-4V, GPT-4o, GTP-4o-mini;

[2]https://platform.openai.com/docs/models

Cohen's kappa	Interpretation
0 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

Table 3
Guidelines for interpreting Cohen's Kappa values.

- Anthropic Models[3]: Claude 3.5 Sonnet, Claude 3 Haiku;

3.6. Prompts

To design the prompts for the models under consideration, we created a template aimed at guiding the models to generate accurate relevance judgments. The only difference lies in the modality setting. In the multimodal setup, where an image is provided, the prompt will reference and include the image. Additionally, we require the model to provide an explanation for its relevance judgment. This element could be useful for interpreting the model's decisions.

Below, we present the prompt used for the Claude family in the text-only scenario. For the complete set of prompts, please refer to the Appendix. In the template, *{{document}}* and *{{query}}* are placeholders for the search result and query, respectively.

Haiku's Prompt Template (Text-only Setup)

```
You are an assistant responsible for rating how the retrieved result is relevant to
the query. Output a token: "2", "1", or "0" followed by a full explanation.
Guidelines:
"2" - The result matches exactly with what the user's query is looking for.
"1" - The result is not exactly with what the user's query is looking for. But it's
pretty similar. As our aim is to be strict on exact matches, this grade is less
likely to be used.
"0" - The result is not related to the query at all.

Result: {{document}}
Query: {{query}}
Output:"
```

It is important to note that these prompt templates are not the result of an extensive exploration of all possible templates. In Section 4.2, we provide a detailed analysis of the prompt engineering process that led to the best performing prompts for Sonnet and Haiku.

4. Results

4.1. Multimodal vs Single-modality Evaluation

The results presented in Table 4 offer several insights into the performance of the considered Large Language Models across different domains and modalities.

[3]https://docs.anthropic.com/en/docs/about-claude/models

	GPT-4v		GPT-4o		GPT-4o mini		Sonnet		Haiku		Human
	MM	Text	MM	Text	MM	Text	MM	Text	MM	Text	MM
Fashion	0.503	0.498	**0.613**	0.606	0.424	0.382	0.441	0.387	0.371	0.431	0.680
Hotel Supplies	0.620	0.596	0.627	0.582	0.506	0.565	0.634	**0.638**	0.471	0.560	0.641
Design	0.320	0.317	**0.404**	0.331	0.294	0.299	0.351	0.381	0.260	0.309	0.447
Average	0.481	0.471	**0.548**	0.506	0.408	0.415	0.475	0.469	0.368	0.433	0.589

Table 4
Cohen's Kappa coefficients between one of the human annotators and the considered Multimodal (MM) models and their text-only (Text) counterpart. Last column shows the inter-annotator agreement among the two human annotators.

Use-case Dependency of LLM Performance The analysis reveals that the LLM performance is dependent on the use case. The models show varying levels of correlation with the human relevance judgements across the different domains. For example, GPT-4v shows higher performance in the Hotel Supplies use case but performs relatively worse in the other areas. We can observe a similar trend across the other models. This variability in model performance is also connected to the inherent difficulty of the tasks. This is also reflected by the varying levels of agreement among the human annotators for the different use cases.

One Model to Rule Them All? The multimodal version of GPT4-o generally performs better than the other models in two out of three cases, achieving the highest average Cohen's Kappa coefficient (0.548). It stands out in the Hotel Supplies and Fashion domains, where it shows substantial agreement with human annotations. However, it is outperformed by Sonnet in the Hotel Supplies domain, suggesting that no single model outperforms all the others across every use case.

Necessity of Multimodal Support In the table we compare each Multimodal (MM) model with their text-only (Text) counterpart. It is worth noticing that the benefits of multimodal support are not uniform across all the models and use cases. For models like GPT-4o, the vision component significantly enhances the performance, increasing the correlation from 0.506 (Text) to 0.548 (MM). This leads to the highest average performance and *remarkably very close to human correlation (i.e. 0.589).* GPT4-v and Sonnet also benefit from the visual component. However, for smaller models, such as Haiku, the vision component appears to have a detrimental effect, decreasing the correlation from 0.433 (Text) to 0.368 (MM). To further investigate the impact of the visual component in Haiku, we performed an ablation study by excluding the textual component and relying solely on the image. Under this configuration, the highest correlation achieved was 0.1 for the Design case, which is significantly lower than the text-only correlation of 0.309. This indicates that for smaller models like Haiku, the visual component may not be sufficiently robust to provide effective multimodal support.

Error Analysis To investigate the poor performance of the smaller multimodal models, we conducted an error analysis on a sample of relevance judgments from Haiku that disagreed with both human annotators. We examined 31 instances of disagreement and identified three distinct error categories. Notably, Haiku generates an explanation for its relevance judgments, which allowed us to categorize the errors effectively.

The most frequent issue (17 cases) involved the model's failure to correctly identify the product type. For example, when given the query "pure cotton dressing gown", the model misclassifies a linen dressing gown, justifying its choice with the explanation: *The product is [...] made of linen, which is a natural fiber similar to cotton.*. In half of the remaining cases, Haiku's errors originates from wrong assumptions. For instance, the model confused the brand name for bras with the word "band" as in "hairband", leading to incorrect judgments. Lastly, 7 of the cases were related to the model's vision capabilities, where it failed to recognize specific patterns or prints on products, resulting in inaccurate relevance assessments. Table 5 provides an example of this type of error, including the image, query, and the explanation generated by Haiku.

Image	Query	Explanation
	h&m kids unicorn printed t-shirt	The provided result does not match the user's query for a "h&m kids unicorn printed t-shirt". The result is for a "Mia l/s top" which is a long-sleeved top in soft, printed cotton jersey, but it does not appear to have a unicorn print. **The image also does not show a unicorn print.** Therefore, the result is not relevant to the user's query, and I would rate it as a "0".

Table 5
Example of an error in Haiku's relevance judgment due to its inability to understand the provided image.

	GPT-4V	GPT-4o	GPT-4o-mini	Sonnet	Haiku
$/1M Input tokens	10.00	5.00	0.15	3.00	0.25
$/1M Output tokens	30.00	15.00	0.60	15.00	1.25
$/1M images (low resolution)	425.00	425.00	425.00	1048.58	87.38

Table 6
Cost comparison across different models for input tokens, output tokens, and image processing. The costs are reported per million units, with image processing based on low-resolution images. Prices reflect the providers' rates as of August 16, 2024.

Cost-Accuray Trade-off Considering the previous results comparing multimodal versus text-only performance, we can make important cost-accuracy trade-off considerations when choosing a model to adopt for relevance judgement. The costs reported in Table 6 refer to the providers' pricing as of August 16 2024. For image processing, calculations are based on handling 1M low-resolution images. Specifically, OpenAI's GPT-4V and GPT-4o allow users to constrain the number of tokens to 85 per image, resulting in a cost of $0.000425 per image. For GPT-4o-mini, this limit is set at 2,833 tokens per image, leading to the same per-image cost. Claude models adjust token usage based on image size, so for fairness, we report their prices for images resized to 512x512 pixels.

Despite its strong performance for both text and multimodal results, GPT-4V is the most expensive LLM with high costs for both input and output tokens and image processing. The more recent GPT-4o, on the other hand, offers higher performance at a lower cost, making it the current best choice when high precision is required. Compared to Claude 3.5 Sonnet, GPT-4o has higher input token costs but benefits from lower image processing costs for images of the same size.

Assuming 800 input tokens per search result (which reflects an estimation of the average number of tokens in our experiments), the cost per result in a text-only setting would be $0.004 for GPT-4o and $0.0024 for Sonnet. In a multimodal setting, these costs would rise slightly to $0.004425 for GPT-4o and $0.00349 for Sonnet, reflecting the additional expense of processing images. Therefore, given that Sonnet is the third-best performing model in terms of correlation with humans, it represents a good choice for scenarios where a moderate budget is available but high-quality results are still important.

For smaller models like GPT-4o-mini and Haiku, the cost differences become significant, at the expense of performance. In a text-only setting, GPT-4o-mini's cost per result is the lowest. The text-only Haiku's cost per result is slightly higher than GPT-4o-mini, but its performance is also higher. Importantly, our experiments revealed that the visual component did not contribute significantly to these smaller models in our use cases. Therefore, the multimodal capabilities of GPT-4o-mini and Haiku should be used with caution, especially considering the high costs associated with image processing—particularly for GPT-4o-mini.

4.2. Prompt Engineering

Motivated by the observation that smaller models do not necessarily require visual inputs to achieve moderate agreement with human judgments, we wondered whether prompt engineering could further close the accuracy gap between these smaller models and the larger LLMs, potentially offering an even better cost-accuracy trade-off. We made the following observations:

Strictness guidelines Many of the initial disagreements with human stemmed from the models being more lenient about the *1* (Okay) grade. Results improved after we appended instructions to prefer grades *2* (Great) and *0* (Bad) – e.g. "As our aim is to be strict on exact matches, this grade is less likely to be used."

Smaller models are more sensitive to prompt length We experimented with progressively making prompts more concise while retaining the completeness of the instructions we were providing. This was particularly helpful with smaller models, ultimately leading to our best performing prompt for Haiku.

Asking for explanations Asking LLMs to provide explanations for the grades helped the model to perform better, and also helped us to understand how to iterate via prompt engineering to make the instructions less ambiguous for the model.

Overall, we were able to meaningfully improve the accuracy of Haiku through prompt engineering (from 0.36 to 0.40). Given that this is not far in accuracy, and 20-40 times cheaper than the GPT family, this makes it a very appealing option for application at large scale. For instance, smaller models can be used to generate larger label sets to explore recall issues, while more expensive models focus on smaller sets to evaluate precision.

Even though we were able to improve model results for this analysis, we could not find a systematic way to reliably optimize model accuracy across the board. As a result, the process of prompt engineering felt more like art than science, and motivates further work to develop systematic ways to discover the upper limits of accuracy for each model size.

5. Conclusion

In this paper, we have presented a new analysis of MLLMs-as-a-Judge, to assess the cost-accuracy trade-offs of relevance judgement capabilities of MLLMs across three multimodal search use cases: Hotel Supplies, Design and Fashion. Various LLMs have shown potential, but no single LLM showed optimal cost-accuracy across all use cases evaluated.

We have found that for any given practitioner looking to choose the best LLM judge for their use case, a comprehensive evaluation of all available models is both time-intensive, financially demanding and require significant amounts of energy, which can have a significant effect on the environment. This motivates future work in the following directions: 1) improving the abilities of general MMLMs across use cases, 2) improving cost and computational efficiency of large MMLMs, and 3) creating small MMLMs that are optimized for judging relevance in cost-optimal ways for more specialized applications.

Acknowledgments

Thanks to the entire Objective team for building many pieces of the puzzle that made this work possible. Special thanks to Lance Hasson, Brian Porter, George Gkotsis, Kuei-da Liao, and Faizaan Merchant.

References

[1] H. A. Rahmani, C. Siro, M. Aliannejadi, N. Craswell, C. L. A. Clarke, G. Faggioli, B. Mitra, P. Thomas, E. Yilmaz, Report on the 1st workshop on large language model for evaluation in information retrieval (llm4eval 2024) at sigir 2024, 2024. URL: https://arxiv.org/abs/2408.05388. arXiv:2408.05388.

[2] D. Chen, R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, L. Sun, Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark, 2024. URL: https://arxiv.org/abs/2402.04788. arXiv:2402.04788.

[3] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, M. Seo, Prometheus: Inducing fine-grained evaluation capability in language models, 2024. URL: https://arxiv.org/abs/2310.08491. arXiv:2310.08491.

[4] L. Zhu, X. Wang, X. Wang, Judgelm: Fine-tuned large language models are scalable judges, 2023. URL: https://arxiv.org/abs/2310.17631. arXiv:2310.17631.

[5] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, 2024. URL: https://arxiv.org/abs/2309.10621. arXiv:2309.10621.

[6] J.-H. Yang, J. Lin, Toward automatic relevance judgment using vision–language models for image–text retrieval evaluation, 2024. URL: https://arxiv.org/abs/2408.01363. arXiv:2408.01363.

[7] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, arXiv preprint arXiv:2402.03216 (2024).

[8] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, Mteb: Massive text embedding benchmark, arXiv preprint arXiv:2210.07316 (2022). URL: https://arxiv.org/abs/2210.07316. doi:10.48550/ARXIV.2210.07316.

A. Prompt Templates

In this section, we report the prompts used for the considered models. In the templates, *{{document}}*, *{{query}}*, and *{{image}}* are placeholders for the search result, query, and image respectively. For the OpenAI's models, the image corresponds to the image URL, while for the Anthropic's models, it corresponds to a base64-encoded image.

Haiku and Sonnet's Prompt Template (Multimodal Setup)

```
You are an assistant responsible for rating how the retrieved result is relevant to the query. If an image
is available, use it to determine the relevance to the query. Output a token: "2", "1", or "0" followed by
a full explanation.
Guidelines:
"2" - The result matches exactly with what the user's query is looking for.
"1" - The result is not exactly with what the user's query is looking for. But it's pretty similar. As our
aim is to be strict on exact matches, this grade is less likely to be used.
"0" - The result is not related to the query at all.

Result: {{document}}
Query: {{query}}
```

```
{{image}}
```

```
Token:
```

GPT4's Prompt Template (Multimodal Setup)

User Role: System

```
You are a helpful assistant designed to output JSON. You are RateGPT, an intelligent assistant that can
score search results based on their relevance to a query and the user's intent behind the query. You should
return JSON with two required fields 'reasoning' and 'score'. In the 'reasoning' field, you can explain
your observations of relevance. When producing a score, use the following grading criteria:
    - 0 (BAD) - Use this grade for a search result if it is not related to user's query at all.
    - 1 (OK) - This grade is for a search result that is not exactly what the user's query is looking for,
but it's pretty similar. As our aim is to be strict on exact matches, this grade is less likely to be used.
    - 2 (GOOD) - The product matches exactly with the user's intent and query. Use this score this if the
search result aligns perfectly with the user's query.

### Query Analysis
Before you start grading, it's essential to understand user's intent by breaking apart the query. Keep in
mind, some queries may be more explicit than others. For instance, if user is searching for a clothing
product, then "Red checkered jacket" is more specific compared to "Red jacket". Another example, if user is
searching for a venue, then "Rock concert in San Francisco this weekend" is more specific compared to "Rock
concert in San Francisco". Therefore, adapt your grading contextually.

Consider all the information from all fields.

Note: All fields should be taken with equal importance. You should adhere strictly to these guidelines
while grading and ensure a holistic evaluation of
the search results based on all considered fields.
```

User Role: User

```
You are given a search query and a search result in json format. If an image is available, use it to
determine the relevance to the query. You must indicate with a score whether the result is relevant or not.
---
Follow the following format.

Query: {{query}}

Result: {{result}}

Score: 0 or 1 or 2

---

Query: {{query}}

Result: {{document}}
```

User Role: User

```
{{image}}
```